

Simulation-Based Optimal Design

PETER MÜLLER
Duke University, USA

SUMMARY

We review simulation based methods in optimal design. Expected utility maximization, i.e., optimal design, is concerned with maximizing an integral expression representing expected utility with respect to some design parameter. Except in special cases neither the maximization nor the integration can be solved analytically and approximations and/or simulation based methods are needed. On one hand the integration problem is easier to solve than the integration appearing in posterior inference problems. This is because the expectation is with respect to the joint distribution of parameters and data, which typically allows efficient random variate generation. On the other hand, the problem is difficult because the integration is embedded in the maximization and has to possibly be evaluated many times for different design parameters.

We discuss four related strategies: prior simulation; smoothing of Monte Carlo simulations; Markov chain Monte Carlo (MCMC) simulation in an augmented probability model; a simulated annealing type approach.

Keywords: DECISION PROBLEMS; EXPECTED UTILITY MAXIMIZATION; STOCHASTIC OPTIMIZATION.

1. INTRODUCTION

Optimal design problems are concerned with maximizing expected utility in a statistical experiment. The maximization is over some design parameter $d \in D$. The experiment is defined by a model $p_d(y | \theta)$, i.e., a distribution of a vector y of observables conditional on some unknown parameter vector θ . The model may depend on the design parameter d , hence the subscript d . The model is completed by a prior distribution $p(\theta)$ for the parameter vector. Utility is a function $u(d, \theta, y)$, for example negative squared error loss $u = -\{\theta - E(\theta | y, d)\}^2$, or something more problem specific like $u =$ total number of successfully treated patients, etc. Since the design parameter d has to be chosen before observing the experiment, we need to maximize the expectation of $u(\cdot)$ with respect to (θ, y) . We can formally state the design problem as

$$d^* = \arg \max_{d \in D} U(d), \text{ where } U(d) = \int u(d, \theta, y) \underbrace{p_d(\theta, y)}_{p(\theta)p_d(y|\theta)} d\theta dy. \quad (1)$$

$U(d)$ is the expected utility for action d . Unless model and likelihood are chosen to allow analytic evaluation of the expected utility integral, the optimal design problem requires numerical solution strategies.

Chaloner and Verdinelli (1995) and Verdinelli (1992) provide an extensive review of analytical and approximate solutions to Bayesian optimal design problems, focusing on the traditional experimental design question of choosing covariates in a regression problem, including non-linear regression and linear regression with interest on a non-linear function of the parameters.

In this paper we explore different, simulation based strategies. Section 2 reviews the basic concept of Monte Carlo simulation for the evaluation of $U(d)$, including a proposal

of "borrowing strength" across simulations under different designs d by smoothing through simulated pairs of design and observed utilities. Section 3 approaches the problem with very different strategies, using a model augmentation which defines an artificial probability model on the triple of design, data and parameters. Simulation in the augmented model is shown to be equivalent to solving the optimal design problem (1). Critical shortcomings of the proposed approach are problems arising with flat and high dimensional expected utility surfaces. Section 4 proposes an idea reminiscent of simulated annealing which replaces the expected utility surface by a more peaked surface without changing the solution of the optimal design problem.

2. PRIOR SIMULATION

Except in special problems, the stochastic optimization problem (1) does not allow a closed form solution. Often the utility function $u(\cdot)$ is chosen to allow analytic evaluation even if a more problem specific utility/loss function were available. A typical example is the use of preposterior variance on some parameters of interest instead of total treatment success in medical decision problems. More realistic utility functions can be used in simulation based solutions to the optimal design problem.

Most simulation based methods for optimal design are based on the observation that the integral in $U(d)$ is easily evaluated by Monte Carlo simulation. In most problems $p_d(\theta, y) = p(\theta)p_d(y|\theta)$ is available for efficient random variate generation, allowing an approximation of $U(d)$ by

$$\hat{U}(d) = \frac{1}{M} \sum_{i=1}^M u(d, \theta_i, y_i), \quad (2)$$

where $\{(\theta_i, y_i), i = 1, \dots, M\}$ is a Monte Carlo sample generated by $\theta_i \sim p(\theta)$, $y_i \sim p_d(y|\theta_i)$. Since the prior probability model typically does not depend on the chosen design d , we do not use a subscript d for $p(\theta)$. But no additional difficulties would arise if the prior were to depend on d in a given application.

Monte Carlo approximations of this type are routinely used in Bayesian optimal design problems. In many examples part of the expected utility integral can be solved analytically, leaving Monte Carlo simulation only for the remaining non-analytic integration. Some recent examples in the literature are the following studies.

Sun, Tsutakawa and Lu (1996) consider optimal design for quantal responses, specifying $U(d)$ as expected posterior variance of some function ϕ of a binary response curve. Amongst other strategies to evaluate $U(d)$, they use a combination of Monte Carlo integration with respect to y and numerical quadrature for integration with respect to θ .

Carlin, Kadane and Gelfand (1998) discuss optimal design in a sequential experiment. The setup allows up to K observations (data monitoring points). The design defines a stopping rule which gives at each time a decision of whether to stop the experiment or continue sampling to the next step. Given that the trial is stopped at a certain stage, the terminal decision problem of choosing one of the two available actions can be solved analytically. The design is parameterized in terms of a vector d of upper and lower bounds which decide continuation of the experiment at each step by specifying continuation if the posterior mean on the parameter of interest falls between the bounds, and termination otherwise. Monte Carlo simulation from $p_d(\theta, y)$ is used to evaluate expected utility for a given design. In the context of the sequential design this strategy is referred to as forward simulation.

A more traditional method to evaluate sequential designs uses backward induction (Chapter 12, DeGroot 1970). Vlachos and Gelfand (1996) use Monte Carlo simulation to evaluate the continuation risk in a backward induction for a sequential design problem.

If at the time of the decision some data $x \sim p(x | \theta)$ are already available then the decision needs to condition on x . The prior $p(\theta)$ is replaced by $p(\theta | x)$. In the evaluation of (2), generating $\theta \sim p(\theta | x)$ possibly requires Markov chain Monte Carlo (MCMC) simulation. For example, Wakefield (1994) considers optimal design strategies in a population pharmacokinetic study. The problem is to find the optimal dosage regimen for a new patient. The decision is made conditional on already observed data x for earlier patients who were administered the same treatment. Wakefield proposes Monte Carlo evaluation of expected utilities using a Monte Carlo average as in (2), replacing the prior $p(\theta)$ by the relevant posterior $p(\theta | x)$. The required posterior sample $\Omega = \{\theta_i, i = 1, \dots, M\}$ is generated by MCMC simulation. Compared to parameter estimation, the solution of the optimal design problem comes at no additional computational cost, since the MCMC posterior sample Ω is already generated. In a similar decision problem, Palmer and Müller (1998) consider the question of choosing an optimal schedule of blood stem cell collections. The decision is a vector d of indicators for each of 10 possible days with $d_i = 1$ if a stem cell collection is scheduled for day i , and $d_i = 0$ otherwise. The design space is the set of all possible 2^{10} vectors of such indicators. We evaluate $U(d)$ by a Monte Carlo approximation like in (2). The decision is made for a new patient, conditional on data x from the previously observed patients. Thus $p(\theta | x)$ replaces $p(\theta)$ as the relevant distribution on θ . As in Wakefield (1994), we proceed by generating via MCMC a sample $\{\theta_i \sim p(\theta | x), i = 1, \dots, M\}$ and evaluating $U(d)$ by (2).

In many problems the expected utility surface $U(d)$ can be assumed continuous. The use of (2) in the context of the optimal design maximization fails to exploit such continuity. Separate large scale Monte Carlo simulation for each new value of d includes inefficient duplication of effort in the following sense. In the course of the simulation, assume we have already evaluated $U(d)$ and consider now another design d' which is close to d (in some appropriate metric). Repeating the full Monte Carlo simulation for $U(d')$ entirely neglects the already available evaluation of $U(d)$.

In Müller and Parmigiani (1996) we propose a numerical optimal Bayesian design scheme which does exploit such continuity if present. First select some designs $d_i \in D$ (possibly on a grid). Then simulate experiments $(\theta_i, y_i) \sim p_{d_i}(\theta, y)$, one for each chosen design. For each simulated experiment (d_i, θ_i, y_i) we evaluate the observed utility $u_i = u(d_i, \theta_i, y_i)$. In a scatterplot of d_i and u_i the integration in (1) can be replaced by a simple scatterplot smoothing $\hat{U}(d)$, and the optimal design can be read off as the mode of the smooth curve. Figure 1 shows two typical examples. Under specific assumptions on the design space, the utility function and the chosen method of scatter plot smoothing we show that the optimal design based on $\hat{U}(d)$ is a consistent estimator for d^* .

Erkanli, Soyer and Angold (1998) use the scheme in a problem concerning the estimation of the prevalence p of a rare disorder in a two phase design. In a first phase an inexpensive screening test is administered. A proportion d_1 of the patients who screened positively are subjected to a more expensive diagnostic test. For the patients who tested negative on the screening test, a proportion d_2 is chosen for the diagnostic test. The design problem is the choice of optimal proportions $d = (d_1, d_2)$ subject to an expected budget constraint. The design criterion is minimum preposterior expected variance for p . We will return to this problem as Example 3 below.

3. AUGMENTED PROBABILITY SIMULATION

In Bielza, Müller and Rios-Insua (1999) and Clyde, Müller and Parmigiani (1995a) we propose to solve the optimal design problem (1) by recasting the problem as a problem of simulation from an augmented probability model $h(d, \theta, y)$. A probability model $h(\cdot)$ is defined in such a

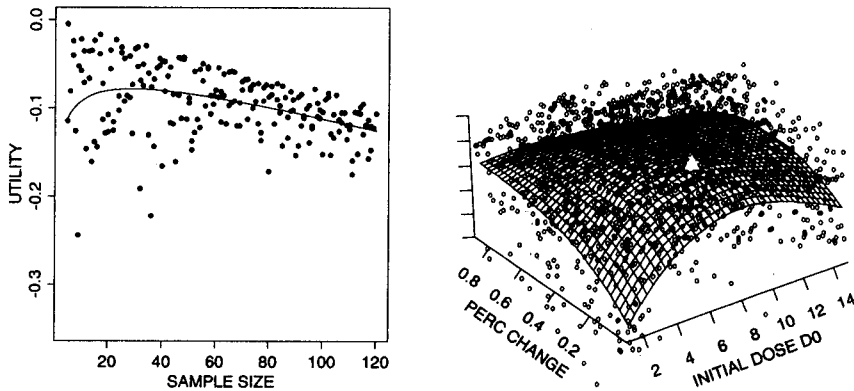


Figure 1. Simulated utilities $u_i = u(d_i, \theta_i, y_i)$ for some designs $d_i, i = 1, \dots, I$. The left panel shows simulated utilities for a design problem with a one-dimensional design parameter ($d = \text{sample size}$ on the horizontal axis). The right panel corresponds to a bivariate design parameter ($d = \{\% \text{ change, initial dose}\}$). In both panels, the points show the observed utilities $u_i = u(d_i, \theta_i, y_i)$ for the simulated experiments (d_i, θ_i, y_i) . The smooth curve/surface is the fitted function $\hat{U}(d)$. Smoothing replaces the integration in (1). The mode of $\hat{U}(d)$ gives the optimal design. In the right panel, the estimated mode is indicated with a white triangle. The left figure shows simulated utilities in Example 1 of Müller and Parmigiani (1996). The example in the right panel is taken from Clyde, Müller and Parmigiani (1995b).

way that the marginal distribution in d is proportional to the expected utility $U(d)$. Specifically, assume D is bounded, $u(d, \theta, y)$ is non-negative and bounded and define an artificial distribution

$$h(d, \theta, y) \propto u(d, \theta, y)p(\theta)p_d(y|\theta)$$

on (d, θ, y) . Under h , the marginal distribution in d is proportional to $\int u(d, \theta, y)p_d(\theta, y)d\theta dy = \bar{U}(d)$, i.e., $h(d)$ is proportional to the expected utility $U(d)$, as desired. The definition of $h(d, \theta, y)$ is reminiscent of data augmentation as used in posterior inference (Tanner and Wong, 1987). But in contrast to data augmentation the marginal distribution of the original random variables (θ, y) is changed in the augmented model. Also, interest in the augmented model focuses on the induced distribution of the artificially added latent variable d , not on the original parameters. Simulation from $h(\cdot)$ can be used to solve the optimal design problem.

Algorithm 1. An MCMC scheme with stationary distribution $h(d, \theta, y)$.

We use super-indices t to indicate design parameters and corresponding utilities after t iterations of the Markov chain.

1. Start with a design d^0 . Simulate (θ, y) from $p_{d^0}(\theta, y) = p(\theta)p_{d^0}(y|\theta)$. Evaluate $u^0 = u(d^0, \theta, y)$.
2. Generate a "candidate" \tilde{d} from a probing distribution $g(\tilde{d}|d^0)$. Details of the probing distribution will be discussed below.
3. Simulate $(\tilde{\theta}, \tilde{y})$, as in 1, under design \tilde{d} . Evaluate $\tilde{u} = u(\tilde{d}, \tilde{\theta}, \tilde{y})$.
4. Compute

$$a = \min \left\{ 1, \frac{h(\tilde{d}, \tilde{\theta}, \tilde{y}) g(d^0|\tilde{d}) p_{d^0}(\theta, y)}{h(d^0, \theta, y) g(\tilde{d}|d^0) p_{\tilde{d}}(\tilde{\theta}, \tilde{y})} \right\} = \min \left\{ 1, \frac{\tilde{u} g(d^0|\tilde{d})}{u^0 g(\tilde{d}|d^0)} \right\}$$

5. Set

$$(d^1, u^1) = \begin{cases} (\tilde{d}, \tilde{u}) & \text{with probability } a \\ (d^0, u^0) & \text{with probability } 1 - a \end{cases}$$

6. Repeat steps 2 through 5 until the chain is judged to have practically converged.

The algorithm defines an MCMC scheme to simulate from $h(d, \theta, y)$, using a Metropolis-Hastings chain with an independence proposal to update (θ, y) . There still remains the specification of the probing distribution $g(\tilde{d} | d)$. Choosing a symmetric probing distribution, that is one for which $g(\tilde{d} | d) = g(d | \tilde{d})$, leads to a simple expression for the acceptance probability, $a = \min(1, \tilde{u}/u)$.

For later reference we note an alternative probing distribution leading to an algorithm which is similar to a Gibbs sampler for posterior simulation. Denote with p the dimension of the design space, and let $d = (d_1, \dots, d_p)$. Replace steps 2 through 5 by a single step 2':

2'. For $j = 1, \dots, p$, generate a new value $d_j^1 \sim h(d_j | d_1^1, \dots, d_{j-1}^1, d_{j+1}^0, \dots, d_p^0)$.

To (approximately) generate from $h(d_j | d_i, i \neq j)$ we use the following random walk Metropolis chain. Proceed like in steps 2 through 5 of Algorithm 1, but instead of $g(\tilde{d} | d)$ use a probing distribution $g_j(\tilde{d} | d)$ which changes only the j -th coordinate, i.e., $\tilde{d}_i = d_i, i \neq j$. Simulating sufficiently many iterations of steps 2 through 5 with g_j as probing distribution will generate an approximate draw from $h(d_j | d_1^1, \dots, d_{j-1}^1, d_{j+1}^0, \dots, d_p^0)$.

Algorithm 1, as well as the above described variation, are special cases of MCMC simulation as described in Tierney (1994). If the design space D is an interval in R^p and $0 \leq u(d, \theta, y) \leq M$ for some bound M , then it follows from results in Tierney (1994) that the above described schemes define Markov chains in (d, θ, y) with stationary distribution h .

The output from the MCMC simulation can be used in several ways to derive the optimal design. First, the sample of simulated d^t can be used to derive an estimate of $h(d)$. The mode corresponds to the optimal design d^* . This is illustrated in the following example.

Example 1. In Bielza *et al.* (1999) we illustrated the scheme in an example taken from Covaliu and Oliver (1995). An electric company has to decide whether to commission a conventional or an advanced reactor. The company has the option of ordering a test of the advanced reactor. The problem involves three decision nodes $d = (d_1, d_2, d_3)$ representing decisions whether or not to commission the test (d_1), and which reactor type to decide for under the possible outcomes of the test, including a dummy outcome corresponding to "no test" (d_2, d_3). There are two random variables $y = (y_1, y_2)$, the outcome of the test (y_1) and the occurrence of an accident (y_2). The utility function includes a cost for commissioning the test and for ordering the two reactor types, and a payoff as a function of possible accidents and reactor types. Decision space and sample space are finite discrete with few possible realizations only, allowing many alternative solution methods including straightforward analytic evaluation of expected utility under all possible decisions. For illustration we solve the problem using simulation on the augmented probability model $h(d, y)$ (there are no unobservable parameters θ in this problem). Figure 2a shows the histogram of simulated values for the decision node d .

In many cases the approach used in Example 1 will be impracticable because the expected utility function, and thus $h(d)$, are too flat around the mode, requiring prohibitive simulation sample sizes to estimate the mode. Instead we consider a scatter plot of all simulated pairs (d, u) and proceed as proposed in the previous section. A smooth fit through the simulated points provides an estimate of $U(d)$, and the optimal design can be readily read off. In this case, the MCMC simulation serves to stir sampling towards areas of high expected utility and replaces sampling on a grid. Note that the pairs (d, u) used for the smoothing have to include *all* simulated designs, including proposals \tilde{d} which were rejected as states of the Markov chain.

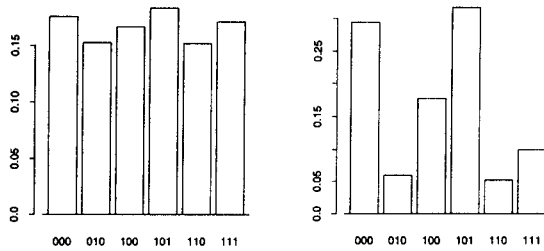


Figure 2. Example 1. The left panel shows $h(d)$ estimated by a histogram of simulated values d^s . The right panel shows the same estimate based on $h_j(d)$ given in (3). Simulation from h_j is discussed in Section 4.

Compared to simulating an equal number of experiments for designs on a regular grid, as proposed in Section 2, the computational effort involved in Algorithm 1 is the same. The evaluation of the acceptance probabilities in step 5 requires no additional computation beyond the evaluation of the observed utilities $u(\cdot)$. Example 2 illustrates this strategy.

Example 2. In Rios-Insua *et al.* (1997), we discuss a problem concerned with the operation of two reservoirs. Four parameters $d = (d_1^K, d_2^K, d_1^C, d_2^C)$ represent the amount of water released through turbines (subscript 1) and spill gates (subscript 2) on two reservoirs (K and C), respectively. A probability model defines a distribution on inflows into the reservoirs. A utility function formally combines a variety of goals related to energy deficit, final storage and water spillage. Using Algorithm 1 we simulated experiments for 100,000 design choices d . Figure 3 shows a smooth fit through the simulated pairs (d, u) . The value in each cell of the grid is the average over all simulated utilities with designs d falling within the respective grid cell.

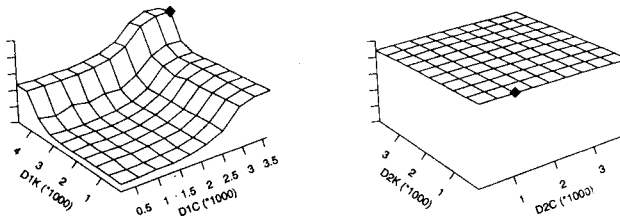


Figure 3. Example 2. The estimated expected utility surface $\hat{U}(d)$. The plots show $\hat{U}(d)$ as a function of (d_1, d_3) and (d_2, d_4) , respectively, keeping the other two coordinates fixed at the estimated optimal value. The solid diamond marks the optimal design.

Using the MCMC simulation incurs no additional computational cost beyond evaluating the utilities $u(d, \theta, y)$ for each considered design. The MCMC simulation keeps sampling focused in areas of high expected utility, and avoids unnecessary sampling in areas of low expected utility. Figure 4 shows a histogram of the number of times that each of 10,000 cells in an equally spaced $10 \times 10 \times 10 \times 10$ grid over the design space were visited. Evenly spreading simulations across the design space, as proposed in Section 2, we would have in all cells an equal number of simulations. Using the MCMC scheme, designs close to d^* get heavily oversampled, as desirable for a more reliable reconstruction of the expected utility surface close to the mode.

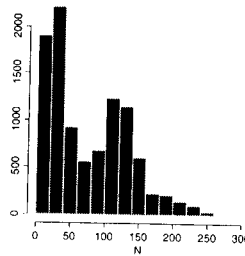


Figure 4. Example 2. A histogram of the number of designs which were considered in each of the 10,000 cells of the grid used in Figure 3. If the designs were evenly spread over the grid all cells would include an equal number of simulations.

The strategy used in Example 2 is feasible only for a low dimensional decision space, say up to around four dimensions. The next section describes a more generally applicable variation of the algorithm.

4. TIGHTENING THE EXPECTED UTILITY SURFACE

Simulation from the augmented model $h(\cdot)$ does not entirely solve the original stochastic optimization problem. It only transforms it to a problem of estimating the mode of a distribution based on a simulated sample. In place of (1) we face the problem of deducing the mode of $h(d)$ from the simulation output. In the last two examples this was possible by considering the histogram of simulated d^i 's, or a surface through all simulated pairs (d, u) . However, simple inspection of a histogram or surface can not provide a general solution for at least two reasons. First, in high dimensional design spaces density estimation becomes impracticable, using histograms or any other method. Similarly, fitting a surface through the simulated (d, u) pairs becomes difficult for high dimensional d . And secondly, expected utility surfaces in most applications are very flat over a wide range of designs. This would require unreasonably large simulation sample sizes to estimate the mode of $h(d)$.

Both problems can be addressed by replacing the target function $h(d)$ with a power transformation $h^J(d)$. This is the generic idea of simulated annealing (van Laarhoven and Aarts, 1987). In the context of simulated annealing the reciprocal $T = 1/J$ is referred to as "annealing temperature". Considering the limit $T \rightarrow 0$ replaces the original target function with a point mass at the mode. For sufficiently large J , simulations from a probability density proportional to the J -th power of the original target function cluster tightly around the mode. Direct application of this scheme to the optimal design problem (1) is hindered by the fact that we do not get to evaluate $h(d)$ itself, but only the augmented model $h(d, \theta, y)$. Taking a power of $h(d, \theta, y)$ or of $u(d, \theta, y)$ would, of course, not achieve the desired transformation of the marginal. However, simulated annealing motivates the following related scheme. Consider

$$h_J(d, \theta_1, y_1, \dots, \theta_J, y_J) \propto \prod_{j=1}^J u(d, \theta_j, y_j) p_d(\theta_j, y_j), \quad (3)$$

i.e., for each d generate J experiments $\{(\theta_j, y_j), j = 1, \dots, J\}$ and consider the product of the observed utilities. The implied marginal in d is proportional to the J -th power of the expected utility, $h_J(d) \propto U^J(d)$. Note that substituting the average instead of the product of the J observed utilities in (3) would imply $U(d)$ as marginal distribution on d , not the desired power $U^J(d)$.

The modification to the earlier MCMC algorithm to generate from h_J is minimal. We replace steps 1, 3 and 4 by:

Algorithm 2. Sampling from $h_J(d)$.

1. Simulate $(\theta_j, y_j) \sim p_{d^0}(\theta, y)$, $j = 1, \dots, J$. For each simulated experiment evaluate $u_j^0 = u(d^0, \theta_j, y_j)$. Define $u^0 = \prod_j u_j^0$.
3. Simulate $(\tilde{\theta}_j, \tilde{y}_j) \sim p_{\tilde{d}}(\theta, y)$, $j = 1, \dots, J$. For each simulated experiment evaluate $\tilde{u}_j = u(d, \tilde{\theta}_j, \tilde{y}_j)$. Define $\tilde{u} = \prod_j \tilde{u}_j$.
4. Compute $a = \min(1, \tilde{u}/u^0)$.

The modified algorithm can be used to replace $U(d)$ by a more peaked transformation $U^J(d)$. This can be done without any notion of an annealing schedule, i.e., using one fixed value J only. For illustration, consider Example 1. Because of the relatively small differences in expected utility, a disproportionately large simulation sample size is needed to detect the optimal design in this simple problem. Using $U^J(d)$ the differences become exacerbated. Figure 2b shows $U^J(d)$, using $J = 10$. The following Example 3 illustrates the same strategy in a more complex decision problem.

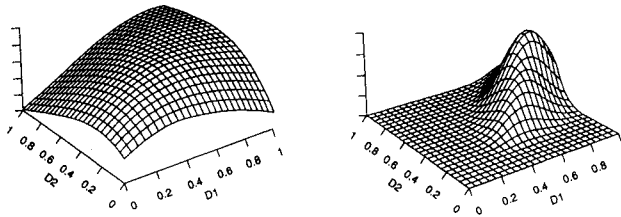


Figure 5. Example 3. The estimated expected utility surface $U(d)$ and the distribution $h_J(d) \propto \{U(d)\}^{20}$ from which designs were simulated. Sampling from h_J focuses simulations around the mode.

Example 3. We implement the modified algorithm for the two phase design problem described in Erkanli, Soyer and Angold (1998). The problem was described earlier in Section 2. Figure 5a shows estimated expected utilities as a function of $d = (d_1, d_2)$. We use the MCMC scheme of Algorithm 2 to select the designs d at which experiments were simulated. This achieves a simulation which concentrates efforts around the mode, and uses only few simulations for areas of the parameter space away from the optimal design. This is illustrated in Figure 5b.

The strategy used in Example 3 relied upon the low dimensionality of the decision vector, and would not be feasible in higher dimensions. For higher dimensional applications Algorithm 2 can be embedded in an annealing scheme. Starting with $J = 1$ use an “annealing schedule” to increase J over iterations, i.e., $J = J(t)$. In Section 3 we described a Gibbs sampler like variation of Algorithm 1. We can define an analogous variation of Algorithm 2. If the decision vector is a discrete p -dimensional vector $d = (d_1, \dots, d_p)$ with $d_j \in \{1, \dots, L\}$, then direct application of Theorem B from Geman and Geman (1984) establishes convergence to the optimal design. Let $U^* = \sup U(d)$, $U_* = \inf U(d)$ and $\Delta = U^* - U_*$. The result requires that the annealing schedule be bounded by $J(t) \leq \log(t)/(p \Delta)$.

For practical use we do not propose a formal implementation of an annealing scheme. Rather, we suggest to increase J only until the simulated designs are sufficiently tightly clustered such that the (trimmed) sample mean of the simulated designs is a reasonable approximation

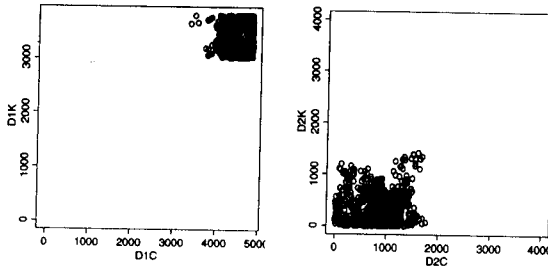


Figure 6. Example 2. Simulation output in a simulated annealing scheme. Note how the simulated points cluster around the mode of expected utility surface shown in Figure 3.

of the optimal design. Figure 6 illustrates this scheme in Example 2. The circles indicate the designs d^t generated in the simulated annealing scheme.

The sample mean of the simulation output provides a good approximation of the optimal design without the intermediate step of reconstructing $h(d)$.

Still, problems where continuity of $U(d)$ is an unreasonable assumption defy this strategy. Designs with high expected utility can not necessarily be expected to cluster together. In such problems we are only left with the option of formal simulated annealing.

Example 4. In Sansó and Müller (1999) we address a problem of choosing rainfall monitoring stations. Out of an existing network of 80 stations a subset of around 40 stations needs to be selected. We define a utility function to formalize the aim of making inference about rainfall over the area on one hand, and minimizing cost on the other hand. The utility function includes a payoff each time the predictive residuals at any of the current 80 locations is in absolute value below a certain threshold δ . The predictive “residual” at a station which is actually included in the subset is defined as zero. The design parameter is a vector $d = (d_1, \dots, d_{80})$ of indicators with $d_i = 1$ if station i is included in the chosen subset, and $d_i = 0$ otherwise. The high dimensionality of the decision vector complicates the design problem. In Sansó and Müller (1999) we used Algorithm 1, together with an idea proposed in Bielza *et al.* (1999). Using a hierarchical clustering tree of the simulated designs we find areas where simulations cluster.

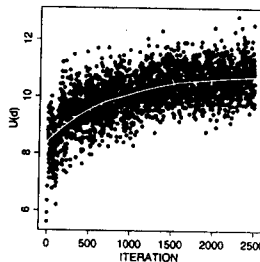


Figure 7. Example 4. Estimated $U(d^t)$ for simulated designs d^t , $t = 1, \dots, T$. Evaluating $U(d^t)$ is not part of the algorithm and was separately computed for this plot, using a Monte Carlo approximation (2) with $M = 100$. The numerical standard errors are approximately 0.5. The white line is a loess fit through the scatterplot showing the steady increase in expected utility.

Alternatively, Figure 7 illustrates the use of the simulated annealing algorithm described above.

5. DISCUSSION

We have proposed several schemes to implement simulation based optimal Bayesian design. None of the proposed algorithms is sufficiently general and robust to allow routine application without adaptation to the specific problem. Rather, the proposed algorithms should be understood as examples for possible strategies.

The discussion was in the context of solving optimal design problems specified in the form of a probability model, a decision space, and a utility function, but could of course be used to solve decision problems defined in any equivalent alternative way. For example, influence diagrams are used for a graphical representation of decision problems. See, for example, Shachter (1986) or Shenoy (1995) for a definition and review of solution strategies. The algorithms defined in this paper can be used as simulation based solution strategies for influence diagrams, including arbitrary, possibly continuous random variables and decision nodes. In fact, Jenzarli (1995) proposed the use of Gibbs sampling estimates (2) to evaluate $U(d)$ in the context of influence diagrams (Gibbs sampling rather than independent sampling is required since the decision is possibly made conditional on some data known at the time of decision making).

An interesting variation in the simulated annealing scheme discussed in Section 4 is the use of probing distributions which incorporate some approximate knowledge of the optimal design. Laud, Berliner and Goel (1992) and Fei and Berliner (1991) discuss such strategies in a more general context.

The optimal design problem (1) is formally similar to the problem of maximum likelihood estimation with normalization described in Geyer (1994). Given a family $\{h_\theta(x) : \theta \in \Theta\}$ of non-negative integrable functions, find $\hat{\theta}$ to maximize the normalized likelihood $l(\theta) = h_\theta(x) / \int h_\theta(x') dx'$. Geyer (1994) discusses a solution based on replacing the integral $c(\theta) = \int h_\theta(x) dx$ by a Monte Carlo estimate based on n Monte Carlo simulations. The paper gives conditions under which the resulting approximation $l_n(\theta)$ hypoconverges to $l(\theta)$, and maximum likelihood estimates $\hat{\theta}_n$ derived from $l_n(\cdot)$ converge to the maximum likelihood estimate $\hat{\theta}$ derived from $l(\cdot)$. Hypoconvergence is a type of convergence of a sequence of functions defined and explained in Geyer (1994).

REFERENCES

- Bielza, C., Müller, P. and Ríos-Insúa, D. (1999). Monte Carlo methods for decision analysis with applications to influence diagrams. *Manag. Sci.* (to appear).
- Carlin, B., Kadane, J. and Gelfand, A. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics* **54**, 964–945.
- Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statist. Sci.* **10**, 273–304.
- Clyde, M., Müller, P., Parmigiani, G. (1995a). Exploring the expected utility surfaces by Markov changes, *Tech. Rep.*, Duke University.
- Clyde, M., Müller, P. and Parmigiani, G. (1995b). Inference and design strategies for a hierarchical logistic regression model. *Bayesian Biostatistics*, (A. D. Berry and D. Stangl, eds.), New York: Marcel Dekker, 297–320.
- Covalliu, Z. and Oliver, R. (1995). Representation and solution of decision problems using sequential decision diagrams. *Manag. Sci.* **41**, 1860–1881.
- DeGroot, M. (1970). *Optimal Statistical Decisions*. New York: McGraw Hill.
- Erkanli, A., Soyer, R. and Angold, A. (1998). Optimal Bayesian two-phase designs for prevalence estimation. *J. Statist. Planning and Inference* **66**, 171–191.
- Fei, L. and Berliner, L. M. (1991). Asymptotic properties of stochastic probing for global optimization: The finite case. *Tech. Rep.* **474**, Ohio State University.
- Geman, S. and Geman A. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intelligence* **6**, 721–740.
- Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations, *J. Roy. Statist. Soc. B* **56**, 261–274.

- Jenzarli, A. (1995). Solving influence diagrams using Gibbs sampling. *Tech. Rep.*, University of Tampa.
- Laarhoven, van P. J. M. and Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Applications*. Dordrecht: Reidel.
- Laud, P. W., Berliner, L. M. and Goel, P. K. (1992). A stochastic probing algorithm for global optimization. *J. of Global Optimization* 2, 209–224.
- Müller, P. and Parmigiani, G. (1996). Optimal design via curve fitting of Monte Carlo experiments. *J. Amer. Statist. Assoc.* 90, 1322–1330.
- Palmer, J. and Müller, P. (1998). Bayesian optimal design in population models of hematologic data. *Statist. in Medicine* 17, 1613–1622.
- Rios-Insua, D., Salewicz, A., Müller, P. and Bielza, C. (1997). Bayesian methods in reservoir operations: the Zambezi river case. *The Practice of Bayesian Analysis*, (S. French, J. Smith, eds.), New York: Wiley, 107–130.
- Sansó, B. and Müller, P. (1999). Redesigning a network of rainfall stations. *Case Studies in Bayesian Statistics IV* (Kass, R., Carlin, B., Carriquiry, A., Catsonis, C., Gelman, A., Verdinelli, I. and West, M., eds.). New York: Springer (to appear).
- Shachter, R. (1986). Evaluating influence diagrams. *Operations Research* 34, 871–882.
- Shenoy, P.P. (1995). A comparison of graphical techniques for decision analysis. *European J. Operations Research* 78, 1–21.
- Sun, D., Tsutakawa, R. and W. Lu, W. (1996). Bayesian design of experiment for quantal responses: What's promised versus what's delivered. *J. Statist. Planning and Inference* 52, 289–306.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* 82, 528–550.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* 22, 1701–1728.
- Verdinelli, I. (1992). Advances in Bayesian experimental design. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 467–448.
- Vlachos, P. and Gelfand, A. (1996). Bayesian decision theoretic design for group sequential medical trials having multivariate patient response. *Tech. Rep.*, University of Connecticut.
- Wakefield, J. (1994). An expected loss approach to the design of dosage regimes via sapling-based methods. *The Statistician* 43, 13–29.

DISCUSSION

MARTIN A. TANNER (*Northwestern University, USA*)

I want to thank Professor Muller for his interesting overview of tools for simulation based optimal design. Clearly these *should be* the tools of choice in designing the experiments of the next millennium.

To help sell these methods to the less enlightened, I feel we need to be more accommodating to those who may need more assistance in transitioning from the horse and buggy to the Starship Enterprise. This assistance runs the gamut from theoretical considerations to practical implementational issues to available software. Thus, my comments are addressed not only to Professor Muller, but to the cohort of colleagues developing and applying these new and exciting tools for optimal design.

The focus of the present paper has been on methods to identify “the” optimal design. To those less grounded in the Art of Bayes, one may wonder what can be said about these designs? For example, in the context of design with no prior data, the θ_i 's are to be drawn from a proper prior $p(\theta)$. How does “the” optimal design depend on the prior? Even in the context where prior data are available and one draws θ_i 's from a posterior, there may still be a dependency (though probably less so) on the prior. A more general question which impacts all approaches to experimental design is the dependency on the model. How does “the” optimal design depend on the model? If the simulation is run assuming model #1, but in fact model #1 is inadequate and model #2 is more appropriate, then there is the classic concern that the optimal design may not allow (or may allow very weakly) for a check on model adequacy. In addition, which

parameters (what features) of the model are well determined by the optimal design and which less so?

Having characterized the operating characteristics of the resulting designs, can we get a better idea of the set-up and performance of the various algorithms to locate this design? The examples considered in this paper are Godzilla-like (and somewhat unfamiliar) in nature. While the methods seem to be able to bring down the beast, one may like to see how the methods work in more familiar (possibly simpler) situations where it is known what one would probably do and where one can see all the various details that go into the "arts and craft" of running these MCMC routines. For example, if the variance on the probing distribution of Section 3 is too small, can one locate the wrong design? In Section 4, in higher dimensions, a cooling schedule does seem to be needed. More information on how to set up and run such a scheme would be appreciated. Can local modes become a serious problem? Can the design at a local model be much different from the design at the global mode? In higher dimensional problems, can one combine MCMC methods with classical response surface methodology to locate a mode?

A foreseeable request by those who design studies would be for (of course, easy-to-use) software. One could imagine a package tailored for a given market (e.g., biostatistics - Phase I/II/III trials) with design implementations for a variety of common situations (e.g., two sample comparisons), important extensions (e.g., censored regression) and some interesting twists (e.g., random effect / hierarchical models). Some users would be interested in the output as a comparison with classical suggestions. Others would be looking for (the only available) help with a difficult problem. Over time, as the brilliance of the newer methods begins to shine through in classically intractable contexts, the less enlightened will become enlightened.

I want to thank Professor Müller for an interesting and stimulating paper.

DONGCHU SUN (*University of Missouri-Columbia, USA*)

It is a pleasure to congratulate Dr. Muller on a very interesting review of promising simulation methods in optimal design. The paper contains a lot of useful ideas.

There is a large literature on optimal experimental design which appeared over the last thirty years. However, most the design work, both classical and Bayesian, has focused on the linear regression model due to its simplicity and wide-spread use. Because of the increasing use of nonlinear models in areas such as biological and clinical sciences, the need for optimal design in nonlinear models has increased in recent years. Nonlinear optimal design problem arise either because a nonlinear function of the parameters in a linear model or the response function itself is nonlinear. A closed form solution for a Bayesian nonlinear optimal design is usually very difficult to obtain. With advances in computer technology and computational tools, the methods for finding optimal designs have switched substantially from mathematical theory to numerical and simulation-based procedures.

The idea of augmented probability simulation puts the problem of optimal design into a framework of Bayesian model selection in the sense that finding the optimal design is essentially a search procedure for finding the best model. While Bayesian model selection has received a great deal of recent attention and is still problem specific, many of the known model selection methods can be used and extended in the context of finding optimal designs. For example, the model of (d, θ, y) for the augmented probability simulation is essentially based on a uniform probability of d . Instead of a constant prior, a proper or noninformative prior on d can be used as long as

$$u(d, \theta, y)p(\theta)p_d(y|\theta)p(d),$$

is integrable with respect to (d, θ, y) . In practice, one often has some idea or prior information on the design d . For example, it is quite common to search the optimal design among these

designs with a fixed number of supporting points (cf. Chaloner and Larntz, 1989). Furthermore, while $u(d, \theta, y)p(\theta)p_d(y|\theta)$ might not be integrable, using prior information on d could solve the problem of integrability.

The utility function $u(d, \theta, y)$ is very general. The utility functions for almost all commonly used design criteria such as D -, c -, and A - optimal designs, do not depend on y , and are certainly special cases here. The paper has assumed boundedness of the design space D and nonnegativity and boundedness of the utility function $u(d, \theta, y)$ for the augmented probability simulation procedure. I understand that these conditions are used for convenience only. In practice, they are rather too strong. For example, if the utility function is the negative squared error loss (i.e., the posterior variance is the design criterion), $u(d, \theta, y) = -\{\theta - E(\theta|y, d)\}^2$, one can certainly add a positive constant so that the utility function is nonnegative and bounded provided that the parameter space is bounded. As in using improper priors, one does not want to have a bounded parameter space. Instead, one hopes that the posterior is proper. One possible solution is to consider the utility function

$$u(d, \theta, y) = \exp[-\{\theta - E(\theta|y, d)\}^2],$$

as it is used in Griffin & Smith (1998). But this may not be ideal since $E(e^W)$ is not the same as $e^{E(W)}$ for a random variable W . Alternatively, one can use the loss function rather than utility function, since the loss function is often nonnegative. Then we may minimize the expected loss function, rather than maximize the expected utility function. If $u(d, \theta, y)p(\theta)p_d(y|\theta)$ is integrable with respect to (d, θ, y) , then $h(d, \theta, y)$ is indeed a proper distribution, even if $u(d, \theta, y)$ is unbounded or the density $p(\theta)$ is improper.

For Bayesian model selection, one often needs to find the marginal likelihood, and methods such as harmonic means from the Gibbs output are sometimes unstable. In using the Laplace-Metropolis estimator of a marginal likelihood, Raftery (1996) suggested using the posterior mean or the multivariate median of the Gibbs sample from the posterior distributions rather than finding posterior modes. It has been noticed that the target function around the optimal design is often very flat in the sense that even if the estimated optimal design is not very accurate, the expected utility function or the expected loss function at the optimal design is still very close to the real maximum. In this case, for the augmented probability simulation, one might just use the means or medians of d based on MCMC output rather than finding the modes of the marginal distribution of d . This is especially useful when the dimension of d is large, design space D is continuous, and it is desirable to search for optimal design for a fixed number of design points.

The case when the design parameter is a vector of indicators is very interesting. For instance, in Example 4 (see also, Sanso and Muller, 1997), the design parameter is a vector $d = (d_1, \dots, d_{80})$, where $d_i = 0$ or 1. Various Bayesian procedures for selecting the best set of predictors for regression can be used to find the best design. I wonder if George and McCulloch's (1993) stochastic search variable selection can be applied efficiently to find the best design among the $2^{80} \approx 1.2 \times 10^{24}$ possible designs.

Simulation methods together with numerical quadrature for integration can be very powerful for the nonregular case when the problem does not fit into a decision framework. For example, Sun and Tsutakawa (1997) considered penalized risk for the design problem. For a fixed design d , prior $p(\theta)$ and loss function $L(a, \theta; d)$, let $\hat{\theta} = \hat{\theta}(y)$ be the Bayesian estimator, which minimizes both the Bayes risk and the posterior risk, given by

$$r_d(\hat{\theta}) = \int L(\hat{\theta}, \theta; d)p_d(y|\theta)p(\theta)dyd\theta \text{ and } r_d(\hat{\theta}|y) = \int L(\hat{\theta}, \theta; d)p_d(\theta|y)d\theta,$$

respectively. Here $p_d(\theta|y)$ is the posterior density of θ given y for a known design d . Sun and Tsutakawa (1997) showed that the optimal design d^* that minimizes $r_d(\hat{\theta})$ with respect to $d \in D$ may not be desirable in the sense that for some observable y , $r_d(\hat{\theta}|y)$ could be very large. It is then proposed to penalize these outcomes whose $r_d(\hat{\theta}|y)$ exceed the Bayes risk. In general, a design criterion with a penalty term could be used, e.g.,

$$H(d) = r_d(\hat{\theta}) + C \int \left[\left\{ r_d(\hat{\theta}|y) - r_d(\hat{\theta}) \right\}_+ \right]^a p_d(y|\theta)p(\theta)dyd\theta,$$

where C and a are two positive constants and $x_+ = \max(0, x)$. A special case for dose response experiments when L is the squared error loss and $C = 1$ or 2 was considered in Sun and Tsutakawa (1997). We found that it is prudent to choose a design which reduces the chances of poor outcomes by giving up a small amount of Bayes risk.

Note that the design criterion $H(d)$ cannot be written into a decision framework. The closed form expression for the optimal design often does not exist. However, a combination of simulation and numerical quadrature make it possible to find the solution easily. For example, suppose that $p_d(y|\theta)p(\theta)$ is integrable with respect to (y, θ, d) and we are able to compute $r_d(\hat{\theta}|y)$ for given (y, d) . We first simulate a random sample $\{(y_i, \theta_i, d_i), i = 1, \dots, m\}$ of (y, θ, d) based on the joint density

$$p(y, \theta, d) \propto p_d(y|\theta)p(\theta).$$

We then fit the response curve based on m points $\{r_{d_i}(\hat{\theta}|y_i), i = 1, \dots, m\}$ on the space D . This curve is an approximation of the Bayes risk $r_d(\hat{\theta})$. Based on the same random sample of (y, θ, d) , we fit another curve of $d \in D$ based on the points

$$\{r_d(\hat{\theta}) + C[\{r_{d_i}(\hat{\theta}|y_i) - r_d(\hat{\theta})\}]^a, i = 1, \dots, m\}.$$

The second curve is then an approximation of $H(d)$, and its minimum is obtained at the optimal d .

R. ALEX REUTTER (*Duke University and SPSS Inc., USA*)

The choice of the value of J (or annealing schedule $J(t)$) seems to be of some importance. A choice of J that is too low leaves the expected utility surface too flat, while a choice of J that is too high causes the generation of many unnecessary experiments.

I would greatly appreciate it if the author can provide references that discuss the annealing schedule in greater detail, or relate experiential evidence for sensible choices for J .

REPLY TO THE DISCUSSION

I thank Professors Sun and Tanner, and Alex Reutter for stimulating discussions. The discussions raise several important issues and pose interesting questions.

Professor Sun's comment about the parallel between optimal design problems and model selection is interesting. From a formal perspective, I agree. Model selection could certainly be set up as a decision problem of the type discussed in the paper. Still, there is an important difference in implicit assumptions. Most methods discussed in this paper exploit some notion of continuity of the expected utility function with respect to the decision parameters. While this is not a strictly required technical condition, it is what makes the methods work in practice. This is especially true for the idea of replacing the expected utility integral by smoothing through a scatterplot of utilities in simulated experiments (Section 2). In contrast, in model selection problems we typically cannot rely on any such notion. Thus I would not expect methods

proposed in this paper to be successful for, e.g., variable selection problems. And, vice versa, I am not hopeful about methods proposed specifically for variable selection to be successful in generic expected utility maximization problems.

Professor Sun suggests considering prior information on the design to improve algorithms. This is a promising idea worth pursuing. In the context of simulated annealing algorithms a similar strategy has been proposed by Fei and Berliner (1991), and could be incorporated in the approach outlined in Section 4.

I agree with Professor Sun's comment that the stated conditions on the utility function, albeit sufficient, are far from necessary and can be significantly relaxed.

Finally, I fully agree with the observation that solving decision problems should go beyond the formal maximization of the expected utility surface. Sometimes it is more important to learn more about the expected utility surface than to exactly pin down the mode. This is especially true when deviations from the formal optimal decision lead to only negligible compromises in expected utility, i.e., the surface is very flat.

This closely relates to Professor Tanner's comments about sensitivity. I entirely agree that sensitivity analysis should be an important part of solving a design problem. This includes sensitivity to the prior model, to the sampling model, and sensitivity to the utility function. While I cannot offer any formal solution to these important problems, I would like to point out one feature of the proposed approaches. Except for the formal implementation of simulated annealing used in Example 4, all proposed algorithms include an estimation of the expected utility surface. Inspection of the estimated surface allows some informal conclusions about sensitivity and robustness. For example, in Figure 3, the surface is very flat in the two decision parameters (d_2, d_4), but very peaked in the other two parameters. This suggests to explore the inclusion of other policy goals related to (d_2, d_4) in the utility function. In practical case studies, the formal utility function summarizes only some of the relevant decision criteria; there are almost always other secondary goals which could be explored.

Professor Tanner requests to see the methods work in simple and conventional design problems. Of course I agree that this is an important exercise to learn more about the algorithms. In Bielza *et al.* (1996) we work through some simple illustrative examples and include a discussion of general issues in comparing the proposed schemes with existing algorithms in the Operations Research literature. In traditional statistical design problems using typical inference loss, like squared error loss, the proposed methods typically do not fare well. There are at least two reasons for this. First, inference loss often requires an additional integration to evaluate the utility $u(d, \theta, y)$ for a simulated experiment. Examples are posterior integrals to find posterior variance, or Kullback-Leibler divergence, etc. Second, traditional design problems often are deliberately set up in such a way that analytical or approximate solutions are possible, leaving simulation based optimal design hopelessly inefficient. The proposed simulation based design algorithms work best in decision problems with loss functions which are easy to evaluate for a realized simulated design, but which defy analytic integration for expected utility evaluations, and which do not offer any obvious approximations.

I share Professor Tanner's interest to see software available to implement optimal design in important application problems. An efficient implementation requires many problem specific details, including in particular a good choice of probing distributions in the MCMC scheme to simulate from the augmented probability model, and any possible partial analytic integration of the expected utility integral. Therefore I would not expect generic implementations to be successful. But implementations for specific classes of problems, as proposed by Professor Tanner, will be an interesting goal to pursue.

Alex Reutter's question how to determine a reasonable value for J addresses an important issue. I have no good answer beyond suggesting simple trial and error. A formal solution to the problem of finding an "optimal" J would defy the purpose because it would likely be a more complex problem than the original design problem itself. Note, however, that there is no "wrong" choice for J either. As a practical guideline I propose to choose J as small as possible, but sufficiently large to narrow down the optimal decision to the desired accuracy. Specifying J too large exacerbates the problem of possibly getting trapped in local modes.

ADDITIONAL REFERENCES IN THE DISCUSSION

- Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Planning and Inference* **21**, 191–208.
- George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.
- Raftery, A. E. (1996). Hypothesis testing and model selection. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 163–187.
- Sun, D. and Tsutakawa, R. K. (1997). Bayesian design for dose responses with extra penalty for unexpected outcomes. *Biometrics* **53**, 1262–1273.